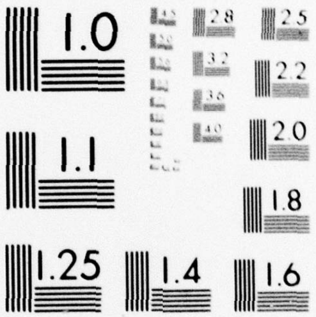
AD-A056 725   STATE UNIV OF NEW YORK AT BUFFALO AMHERST   STATISTICA--ETC   F/G 12/1
MAXIMUM ROBUST LIKELIHOOD ESTIMATION.(U)
JUL 78   E PARZEN   DAAG29-76-G-0239
UNCLASSIFIED   ARO-13845.2-M   NL

1 OF 1
AD A056 725

END
DATE
FILMED
9-78
DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER (14) 13845.2-M	2. GOVT ACCESSION NO. (18) ARO
3. RECIPIENT'S CATALOG NUMBER	(12)
4. TITLE (and Subtitle) (6) Maximum Robust Likelihood Estimation.	5. TYPE OF REPORT & PERIOD COVERED (9) Final Report, 5 Jun 76 - 31 Jul 78
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) Emanuel Parzen	8. CONTRACT OR GRANT NUMBER(s) (15) DAAG29-76-G-0239
9. PERFORMING ORGANIZATION NAME AND ADDRESS State University of New York, Buffalo Amherst, New York 14226	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P. O. Box 12211 Research Triangle Park, NC 27709	12. REPORT DATE (11) July 1978 13. NUMBER OF PAGES 10
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) (12) 44 p.	15. SECURITY CLASS. (of this report) unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

After a statement describing the over-all goals and personnel of this research project, this final report contains: a list of technical reports on research supported by this project, and a description of research accomplishments as given in the abstracts or introductions of the technical reports which have been issued.

409 544

78 07 21 031

MAXIMUM ROBUST LIKELIHOOD ESTIMATION

Final Report of a Research Project
June 1976 - July 1978

Directed by

EMANUEL PARZEN

at Statistical Science Division
State University of New York at Buffalo

Supported by

ARMY RESEARCH OFFICE
Grant DA AG29-76-0239

Introduction

After a statement describing the over-all goals and personnel of this research project, this final report contains: a list of technical reports on research supported by this project, and a description of research accomplishments as given in the abstracts or introductions of the technical reports which have been issued.

Goals

This research has developed a general approach to statistical data analysis (in particular to non-parametric statistical data modeling and to robust analysis and modeling of statistical data, including the one-sample, two-sample, bivariate-sample and multivariate-sample cases).

The new results being obtained seem to be attracting wide interest: (1) Professor Parzen's paper "Nonparametric Statistical Data Modeling" is a major invited address at the August 1978 Annual Meeting of the American Statistical Association and will be published with discussion in the December 1978 issue of the Journal of the American Statistical Association; (2) Professor Parzen's paper "A Density-Quantile Function Perspective on Robust Estimation" was given at the April 1978 ARO Symposium on Robust Estimation and will be published in its proceedings.

Personnel

The following faculty have worked on this research project:

Emanuel Parzen
Marcello Pagano
H. Joseph Newton
Jean-Pierre Carmichael

The following Ph. D. students have worked on this research project:

David Trichtler
Michael White

Technical Reports

Author	Title	Date and Report No.
Emanuel Parzen	Nonparametric Statistical Data Science: A Unified Approach Based on Density Estimation and Testing for "White Noise"	January 1977 47, ARO-1
Jean-Pierre Carmichael, Emanuel Parzen	New Nonparametric Approach to the Two Sample Problem	July 1977 56, ARO-2
Emanuel Parzen	Nonparametric Statistical Data Modeling	January 1978 59, ARO-3
Emanuel Parzen	A Density-Quantile Function Perspective on Robust Estimation	March 1978 60, ARO-4
Jean-Pierre Carmichael	Techniques of Quantile Regression	June 1978 ARO-5

A publication on research originating from support by this ARO project is: Marcello Pagano (1977), "An Approach to Time Series Prediction." Proceedings of the Computer Science and Statistics Tenth Annual Symposium on the Interface.

NONPARAMETRIC STATISTICAL DATA SCIENCE: A UNIFIED APPROACH BASED ON DENSITY ESTIMATION AND TESTING FOR "WHITE NOISE"

by

Emanuel Parzen

The aim of this paper is to introduce a single canonical problem to which one can transform many basic statistical inference and statistical data analysis problems. This canonical problem is most simply described as the problem of testing for white noise via density estimation or smoothing. We first state some of the inference problems which we seek to unify.

One-sample (univariate) inference problems. Let $X_1, \ldots, X_n$ be i.i.d. (independent identically distributed) random variables with common a.c. (absolutely continuous) d.f. (distribution function) $F(x)$ and probability density function $f(x)$. One seeks to efficiently:

(i) estimate $f(x)$ non-parametrically (without making any prior assumption about its functional form)

(ii) test for a specified probability density $f_0(x)$ whether there exists constants μ and σ such that

$$f(x) = \frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right) , \quad F(x) = F_0\left(\frac{x-\mu}{\sigma}\right) .$$

(iii) estimate the parameters μ and σ (called location and scale parameters).

Two-sample (univariate) inference problems. Let $X_1, \ldots, X_m$ be i.i.d. with common a.c. d.f. $F(x)$ and let $Y_1, \ldots, Y_n$ be i.i.d. with common a.c. d.f. $G(x)$. One seeks to efficiently:

(i) test whether there exists constants μ and σ such that

$$G(x) = F\left(\frac{x - \mu}{\sigma}\right) \quad ;$$

(ii) estimate μ and σ .

One-sample multivariate inference problems. Let

$$\underline{X} = \begin{pmatrix} X_1 \\ \cdots \\ X_d \end{pmatrix}$$

be a random vector with absolutely continuous multivariate distribution function $F(x_1, \ldots, x_d)$ and density $f(x_1, \ldots, x_d)$; let $\underline{X}_1, \ldots, \underline{X}_n$ be a random sample. One seeks to efficiently:

(i) test whether the components $X_1, \ldots, X_d$ are independent random variables,

(ii) estimate the multivariate density f ,

(iii) estimate the regression function

$$\mu(x_1, \ldots, x_{d-1}) = E[X_d | X_1 = x_1, \ldots, X_{d-1} = x_{d-1}] \; .$$

In addition, there are multi-sample univariate inference problems and multi-sample multivariate inference problems concerned with the equality of many distributions; however, they are not discussed in this paper.

NEW NONPARAMETRIC APPROACH TO THE TWO-SAMPLE PROBLEM

by

Jean-Pierre Carmichael

and

Emanuel Parzen

ABSTRACT

Given two random samples $(X_1, \ldots, X_m)$ and $(Y_1, \ldots, Y_n)$, we want to test the hypothesis that $F_X(\cdot) = F_Y(\cdot)$. There are different possible alternatives. Here we are mostly concerned about change of location:

$$F_Y(x) = F_X(x - \mu) \quad .$$

In Chapter 1, we review the classical parametric and non-parametric procedures that are currently used. In Chapter 2, we introduce some new test statistics obtained from Parzen's new formulation of the problem (1977). In Chapter 3, we present the results of simulations comparing these different procedures on a wide range of underlying distributions. In Chapter 4, we document the use of a computer package developed here, including some new graphical displays.

NONPARAMETRIC STATISTICAL DATA MODELING

by

Emanuel Parzen

Introduction

It is the aim of this paper to introduce new types of keys for exploratory data analysis (of continuous data) based on estimating the quantile function and density quantile function. It appears that this approach leads to an exploratory data analysis which has a firm probability base. Consequently the distinction between exploratory and confirmatory data analysis can be regarded as a distinction between confirmatory non-parametric statistical data analysis or modeling, and confirmatory parametric statistical data analysis.

The basic proposition of this paper is that exploratory data analysis and conventional parametric statistical inference both have as their aim the estimation of the quantile function $Q(u)$, $0 \leq u \leq 1$, of a random variable X of which the data $X_1, \ldots, X_n$ are independent (or dependent) observations. To estimate Q, one assumes a representation for it of the form

$$Q(u) = \mu + \sigma Q_0(u) \quad .$$

which is equivalent to the classic location and scale parameter model for the probability density function: $f(x) = \frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right)$. We call this representation hypothesis H_0 . One can distinguish four stages of this model.

I. Parametric model: one assumes Q_0 known. Then one's aim is to estimate μ and σ . One uses either maximum likelihood estimation or optimal linear combinations of order statistics.

II. Goodness of fit: one tests H_0 for various specifications of Q_0 (corresponding to the familiar probability laws, such as normal, exponential, logistic, Weibull, Pareto, Cauchy, and so on).

III. Robust parametric model: Q_0 is specified by specifications which permit small deviations from an Ideal Model, such as "Q_0 symmetric and possibly long tailed" or "Q_0 normal except for contamination by outliers."

IV. Non-parametric model: estimate Q_0 , either by estimating the density quantile function $fQ(u) = f\left(Q(u)\right)$, or through suitable plots of the sample quantile functions of transformations of the data.

The main aim of this paper is to introduce a "density estimation" approach to Goodness of Fit tests which also yields estimations of Q . To a specified hypothesis H_0 ; $Q(u) = \mu + \sigma Q_0(u)$, one can define a density $d(u)$, $0 \leq u \leq 1$, such that H_0 is equivalent to $d(u) \equiv 1$. Estimation of $d(u)$ provides a test of H_0 and also an estimator of the true fQ function when H_0 is rejected. Many density estimation methods are available; we believe the "autoregressive" method works best for small samples, and we describe it in detail.

A DENSITY-QUANTILE FUNCTION PERSPECTIVE ON ROBUST ESTIMATION

by

Emanuel Parzen

A perspective on robust estimation is discussed, with three broad sets of conclusions. Point I: The means must be justified by the ends. Point II: Graphical goodness-of-fit procedures should be applied to data to check if they are adequately fitted by the qualitatively defined models which are implicitly assumed by robust estimation procedures. Point III: There is a danger that researchers may regard robust regression procedures as a routine solution to the problem of modeling relations between variables without first studying the distribution of each variable.

New tools introduced include: Student's window; Quantile-Box Plots; density-quantile estimation approach to goodness-of-fit tests; and a definition of statistics as "arithmetic done by the method of Lebesgue integration."

TECHNIQUES OF QUANTILE REGRESSION

by

Jean-Pierre Carmichael

Introduction

Given observations $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ on random variables (X, Y) with joint distribution $F_{X,Y}(x, y)$, we want to estimate the regression function of Y on X , $E[Y \mid X = x]$, nonparametrically.

In order to find a natural estimator (simple computationally and intuitively appealing), Parzen (1977) developed the following theoretical approach.

1. Theoretical Approach:

Let $U_1 = F_X(X)$ and $U_2 = F_Y(Y)$, then the joint distribution of U_1 and U_2 is

$$D_{U_1, U_2}(u_1, u_2) = F_{X,Y}\left(Q_X(u_1),\ Q_Y(u_2)\right)$$

and their joint density is

$$d_{U_1, U_2}(u_1, u_2) = \frac{f_{X,Y}\left(Q_X(u_1),\ Q_Y(u_2)\right)}{f_X\left(Q_X(\mu_1)\ \ f_Y\ Q_Y(u_2)\right)}$$

where F_Z is the distribution function of Z

f_Z is its density function

Q_Z is its quantile function

Let $r(x)$ be the regression function of Y on $X = x$.

$$r(x) = E[Y \mid X = x] = \int_{-\infty}^{\infty} \frac{y\, f_{X,\,Y}(x, y)\, dy}{f_X(x)}$$

We now define the regression-quantile function $rQ\,(\cdot)$ by

$$rQ(u) = r\Big(Q_X(u)\Big) = E[Y \mid X = Q_X(u)]$$

How do we compute $rQ(\cdot)$?

By definition,

$$rQ(u) = \int_{-\infty}^{\infty} \frac{y\, f_{X,\,Y}\Big(Q_X(u), y\ dy\Big)}{f_X\Big(Q_X(u)\Big)}$$

Let $y = Q_Y(u_2)$, then

$$rQ(u) = \int_0^1 Q_Y(u_2)\ d_{U_1,\,U_2}(u, u_2)\ du_2$$

If we introduce a Dirac delta function, we can express $rQ(\cdot)$ as a double integral

1.1 $$rQ(u) = \int_0^1 \int_0^1 Q_Y(u_2)\ \delta(u_1 - u)\ d\ D_{U_1,\,U_2}(u_1, u_2)$$

We estimate $rQ(\cdot)$ by

1.2 $$\hat{r}Q(u) = \int_0^1 \int_0^1 \hat{Q}_Y(u_2)\ \frac{1}{h(n)}\ K\ \left(\frac{u_1 - u}{h(n)}\right) d\ \hat{D}_{U_1,\,U_2}(u_1, u_2)\quad .$$